

# An optimal $Q$ -state neural network using mutual information

D Bollé and T Verbeiren

Instituut voor Theoretische Fysica, KU Leuven, B-3001 Leuven, Belgium

E-mail: [desire.bolle@fys.kuleuven.ac.be](mailto:desire.bolle@fys.kuleuven.ac.be),  
[toni.verbeiren@fys.kuleuven.ac.be](mailto:toni.verbeiren@fys.kuleuven.ac.be)

**Abstract.** Starting from the mutual information we present a method in order to find a hamiltonian for a fully connected neural network model with an arbitrary, finite number of neuron states,  $Q$ . For small initial correlations between the neurons and the patterns it leads to optimal retrieval performance. For binary neurons,  $Q = 2$ , and biased patterns we recover the Hopfield model. For three-state neurons,  $Q = 3$ , we find back the recently introduced Blume-Emery-Griffiths network hamiltonian. We derive its phase diagram and compare it with those of related three-state models. We find that the retrieval region is the largest.

PACS numbers: 87.18.Sn, 05.20.-y, 87.10.+e

One of the challenging problems in the statistical mechanics approach to associative memory neural networks is the choice of the hamiltonian and/or learning rule leading to the best retrieval properties including, e.g., the largest retrieval overlap, loading capacity, basin of attraction, convergence time. Recently, it has been shown that the mutual information is the most appropriate concept to measure the retrieval quality, especially for sparsely coded networks but also in general ([1, 2] and references therein).

A natural question is then whether one could use the mutual information in a systematic way to determine a priori an optimal hamiltonian guaranteeing the properties described above for an arbitrary scalar valued neuron (spin) model. Optimal means especially that although the network might start initially *far* from the embedded pattern it is still able to retrieve it.

In the following we answer this question by presenting a general scheme in order to express the mutual information as a function of the relevant macroscopic parameters like, e.g., overlap with the embedded patterns, activity, ... and constructing a hamiltonian from it for general  $Q$ -state neural networks. For  $Q = 2$ , we find back the Hopfield model for biased patterns [3] ensuring that this hamiltonian is optimal in the sense described above. For  $Q = 3$ , we obtain a Blume-Emery-Griffiths type hamiltonian confirming the result found in [4]. However, in that paper the properties of this hamiltonian have not been discussed, rather the dynamics for an extremely diluted version of the model has been treated. Hence, we derive the thermodynamic phase diagram for the fully connected network modeled by this hamiltonian and show, e.g., that it has the largest retrieval region compared with the other three-state models known in the literature.

Consider a network of  $N$  neurons  $\Sigma_i$ ,  $i = 1, \dots, N$ , taking different values,  $\sigma_i$ , from a discrete set of  $Q$  states,  $\mathcal{S}$ , with a certain probability distribution. In this network we want to store  $p = \alpha N$  patterns  $\Xi_i^\mu$ ,  $\mu = 1, \dots, p$ , taking different values,  $\xi_i^\mu$ , out of the same set  $\mathcal{S}$  with a certain probability distribution. Both sets of random variables are chosen to be independent identically distributed with respect to  $i$

We want to study the mutual information between the neurons and the patterns, a measure of the correlations between them. At this point we note that, since the interactions are of infinite range, the neural network system is mean-field such that the probability distributions of all the neurons and all the patterns are of product type, e.g.,  $p(\{\sigma_i\}) = \prod_i p(\sigma_i)$ . Furthermore, in a statistical mechanical treatment any order parameter  $O^\mu$ , being a function of the neurons and the patterns, can be written in the thermodynamic limit  $N \rightarrow \infty$ , as

$$O^\mu = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N O(\sigma_i, \xi_i^\mu) = \sum_{\sigma \in \mathcal{S}} \sum_{\xi \in \mathcal{S}^p} p_{\Sigma\Xi}(\sigma, \xi) O(\sigma, \xi^\mu) \quad (1)$$

where the left hand side is the configurational average,  $\xi = \{\xi^\mu\}$  and where  $p_{\Sigma\Xi}(\sigma, \xi)$  is the joint probability distribution of the neurons and the patterns. Hence, we can forget about the index  $i$  in the sequel.

The mutual information  $I$  for the random variables  $\Sigma$  and  $\Xi$  is then given by (see, e.g., [5])

$$I(\Sigma, \Xi) = \sum_{\sigma \in \mathcal{S}} \sum_{\xi \in \mathcal{S}^p} p_{\Sigma\Xi}(\sigma, \xi) \ln \left( \frac{p_{\Sigma\Xi}(\sigma, \xi)}{p_{\Sigma}(\sigma) p_{\Xi}(\xi)} \right). \quad (2)$$

A good network would be one that starts initially *far* from a pattern but is still able to retrieve it. Far in this context means that the random variables, neurons and patterns, are almost independent. When  $\Sigma$  and  $\Xi$  are completely independent, then  $p_{\Sigma\Xi} = p_{\Sigma}p_{\Xi}$ . Consequently, when they are almost independent we can write

$$p_{\Sigma\Xi} = p_{\Sigma}p_{\Xi} + \Delta_{\Sigma\Xi} \quad (3)$$

with  $\Delta_{\Sigma\Xi}$  small pointwise. We remark that

$$\sum_{\sigma, \xi} \Delta_{\Sigma\Xi}(\sigma, \xi) = 0. \quad (4)$$

Plugging the relation (3) into the definition (2) and expanding the logarithm up to second order in the small correlations,  $\Delta$ , we find using (4)

$$I(\Sigma, \Xi) = \frac{1}{2} \sum_{\sigma, \xi} \frac{(\Delta_{\Sigma\Xi}(\sigma, \xi))^2}{p_{\Sigma}(\sigma)p_{\Xi}(\xi)} + O(\Delta^3) = \frac{1}{2} \langle \langle (\Delta_{\Sigma\Xi}(\sigma, \xi))^2 \rangle_{\sigma} \rangle_{\xi} + O(\Delta^3) \quad (5)$$

with obvious notation. This approximation is in fact very natural. It is the average over the square of the difference between the correlated and uncorrelated probability distribution.

We remark that still all the patterns are contained in (5). Without loss of generality we consider only one condensed patterns and omit the index  $\mu$  in the sequel. Consequently, only first and second order correlations of the variables will be used, higher order ones can be neglected.

Next, we want to express  $\Delta_{\Sigma\Xi}$  in terms of macroscopic, physical quantities of the system (order parameters). Referring to (1) we write down the following  $Q^2$  moments

$$m^{cd} = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_i \sigma_i^c \xi_i^d = \langle \sigma^c \xi^d \rangle_{\{\sigma, \xi\}} = \sum_{\sigma, \xi} p_{\Sigma\Xi}(\sigma, \xi) \sigma^c \xi^d \quad (6)$$

with  $c, d = 0, \dots, Q-1$  and using the notation  $0^0 = 1$ . We remark that  $m^{00} = 1$  such that we have in general  $Q^2 - 1$  independent parameters specifying  $p_{\Sigma\Xi}(\sigma, \xi)$ .

Up to now the derivation is valid for general  $Q$ -state scalar-valued neurons. To fix the ideas we choose the neuron states as

$$\sigma_c = -1 + \frac{c-1}{Q-1} \quad \text{with } c = 1, \dots, Q. \quad (7)$$

This choice corresponds to a  $Q$ -state Ising-type architecture leading to

$$m^{cd} = \sum_{x, y=1}^Q T^{cdxy} p_{\Sigma\Xi}(\sigma_x, \xi_y), \quad \text{with } T^{cdxy} = \left( -1 + \frac{x-1}{Q-1} \right)^c \left( -1 + \frac{y-1}{Q-1} \right)^d. \quad (8)$$

In a similar way we introduce  $A$  by

$$m^{c0} = \sum_x A^{cx} p_{\Sigma}(\sigma_x), \quad m^{0c} = \sum_x A^{cx} p_{\Xi}(\xi_x), \quad A^{cx} = \left( -1 + \frac{x-1}{Q-1} \right)^c. \quad (9)$$

Finally, by introducing the inverse transformations  $S$  and  $B$

$$\sum_{x,y} T^{cdxy} S_{xyc'd'} = \delta_{a,a'} \delta_{b,b'}, \quad \sum_x A^{cx} B_{xd} = \delta_{c,d}, \quad (10)$$

we can write the approximation of the mutual information up to order  $\Delta^2$  as

$$\tilde{I} = \frac{1}{2} \sum_{x,y} \frac{(\sum_{cd} S_{xycd} m^{cd} - \sum_{cd} B_{xc} B_{yd} m^{c0} m^{0d})^2}{\sum_{cd} B_{xc} B_{yd} m^{c0} m^{0d}} \quad (11)$$

where we have left out the dependence on  $\Sigma$  and  $\Xi$ .

Using (1), the expression (11) for  $\tilde{I}$  can be written in terms of configurational averages of the system. In this way, for large  $N$ , we can express  $\tilde{I}$  as a function of the microscopic variables  $\sigma_i$  and  $\xi_i$ . Using (11) for every pattern  $\mu$ , summing over  $\mu$  and multiplying by  $N$  we get an extensive quantity, denoted by  $\tilde{I}_N$ , which grows monotonically as a function of the correlation between spins and patterns. Therefore,  $H = -\tilde{I}_N$  is a good candidate for a hamiltonian.

Configurational averages also enter into the denominator of (11). Since we are mainly interested in the correlations between spins and patterns, rather than in the respective single probability distributions we assume that the latter are equal such that we use the known distribution of the patterns in the denominator.

What we have presented up to now is a scheme to calculate a hamiltonian for a general  $Q$ -state network with an Ising-type architecture using mutual information.

Next, we discuss some specific examples,  $Q = 2$  and  $Q = 3$ , in detail. We start with  $Q = 2$  states. Given the probabilities associated with each state, the inversion of the transformations  $T$  (8) and  $A$  (9) leads to

$$p_{\Sigma\Xi}(\sigma, \xi) = \frac{1 - m^{10} - m^{01} + m^{11}}{4} \delta_{\sigma,-1} \delta_{\xi,-1} + \frac{1 - m^{10} + m^{01} - m^{11}}{4} \delta_{\sigma,-1} \delta_{\xi,1} \\ + \frac{1 + m^{10} - m^{01} - m^{11}}{4} \delta_{\sigma,1} \delta_{\xi,-1} + \frac{1 + m^{10} + m^{01} + m^{11}}{4} \delta_{\sigma,1} \delta_{\xi,1}. \quad (12)$$

The distributions  $p_{\Sigma}(\sigma)$  and  $p_{\Xi}(\xi)$  can be found by summing out  $\xi$  and  $\sigma$  respectively. Using these distributions,  $\tilde{I}$  becomes

$$\tilde{I} = \frac{(m^{11} - m^{01} m^{10})^2}{2(1 - (m^{01})^2)(1 - (m^{10})^2)}. \quad (13)$$

Substituting the averages over the probability distributions by configurational averages and putting  $m^{10} = m^{01} = b$  in the denominator, where  $b$  is the bias of the patterns, we get

$$H = -\frac{1}{2(1 - b^2)^2} N \sum_{\mu} \left( \frac{1}{N} \sum_i \sigma_i \xi_i^{\mu} - b \sigma_i \right)^2. \quad (14)$$

This hamiltonian can be written as

$$H = -\frac{1}{2} \sum_{ij} J_{ij} \sigma_i \sigma_j \quad \text{with} \quad J_{ij} = \frac{1}{N(1 - b)^2} \sum_{\mu} (\xi_i^{\mu} - b)(\xi_j^{\mu} - b) \quad (15)$$

and this is precisely the Hopfield hamiltonian with [3] and without [6] bias.

We remark that a particularly nice aspect of this treatment is that the adjustment of the learning rule due to the bias enters in a natural way. Furthermore, we learn that the Hopfield hamiltonian is the optimal two-state hamiltonian in the sense that we started from the mutual information calculated for an initial state having a small overlap with the condensed pattern. This confirms a well-known fact in the literature.

One could ask what happens when one assumes that initially the state of the network is already *close* to the embedded pattern. Since the mutual information for fully correlated random variables is equal to the entropy,  $S(\Xi)$ , [5] one is interested in (assuming again one condensed pattern)  $F(\Sigma, \Xi) = I(\Sigma, \Xi) - S(\Xi)$ . We define

$$p_{\Sigma\Xi}(\sigma, \xi) = \sum_{\sigma', \xi'} [p_{\Sigma\Xi}^d(\sigma', \xi') \delta_{\sigma', \sigma} \delta_{\xi', \xi} + p_{\Sigma\Xi}^{od}(\sigma', \xi') (1 - \delta_{\sigma', \sigma} \delta_{\xi', \xi})] \quad (16)$$

with obvious notation. Writing

$$p_{\Sigma\Xi}^d = p_{\Sigma} + \Delta'_{\Sigma\Xi} \quad (17)$$

with  $\Delta'_{\Sigma\Xi}$  small pointwise for large correlations and assuming that  $p_{\Sigma\Xi}^{od}(\sigma, \xi) = 0$  for  $\forall \sigma, \xi$ , in order to retain only the polynomial behaviour, we expand  $F$  and find

$$F(\Sigma, \Xi) = \frac{1}{2} \sum_{\sigma, \xi} \frac{(\Delta'_{\Sigma\Xi}(\sigma, \xi))^2}{p_{\Sigma}(\sigma)} + O(\Delta'^3). \quad (18)$$

Expressing  $F$  in terms of the order parameters as in (11), we get the hamiltonian [8, 9]

$$H = N(1 - b^2) \prod_{\mu} (1 - m_{\mu}^2) \quad \text{with} \quad m_{\mu} = \frac{1}{N} \sum_i \frac{(\xi_i^{\mu} - b)\sigma_i}{1 - b^2} \quad (19)$$

for one pattern. In [9] it is shown that this hamiltonian can store an infinite number of patterns. This is consistent with the intuitive idea that it is possible to store a lot of patterns as long as the network state is initially close to them.

For  $Q = 3$  we focus, without loss of generality, on the case where the distributions are taken symmetric around zero, meaning that all the odd moments vanish. Following the scheme proposed above we arrive at

$$\tilde{I} = \frac{1}{2} \frac{1}{m^{02} m^{20}} (m^{11})^2 + \frac{1}{2} \frac{1}{m^{02} m^{20} (1 - m^{02}) (1 - m^{20})} (m^{22} - m^{02} m^{20})^2. \quad (20)$$

Identifying  $m^{02} = a$  as the activity of the patterns,  $m^{20} = q$  as the activity of the neurons,  $m^{11} = m$  as the overlap,  $m^{22} = n$  as the activity overlap [2], and defining  $l = n - aq$  we arrive at

$$\tilde{I} = \frac{1}{2} \frac{1}{a^2} m^2 + \frac{1}{2} \frac{1}{(a(1 - a))^2} l^2. \quad (21)$$

This leads to a hamiltonian

$$H = -\frac{1}{2} \sum_{i,j} J_{ij} \sigma_i \sigma_j - \frac{1}{2} \sum_{i,j} K_{ij} \sigma_i^2 \sigma_j^2 \quad (22)$$

with

$$J_{ij} = \frac{1}{a^2 N} \sum_{\mu=1}^p \xi_i^{\mu} \xi_j^{\mu}, \quad K_{ij} = \frac{1}{(a(1 - a))^2 N} \sum_{\mu=1}^p \eta_i^{\mu} \eta_j^{\mu}, \quad \eta_i^{\mu} = (\xi_i^{\mu})^2 - a. \quad (23)$$

This hamiltonian resembles the Blume-Emery-Griffiths (BEG) hamiltonian [10]. The derivation above confirms the result found in [4] starting from an explicit form of the mutual information for  $Q = 3$ . In that paper the dynamics has been studied for an extremely diluted asymmetric version of this model. Here we want to discuss the fully connected architecture and derive the thermodynamic phase diagram, which has not been done in the literature, in order to compare it with the other  $Q = 3$  state models known.

In order to calculate the free energy we use the standard replica method [11]. Starting from the replicated partition function and assuming replica symmetry we obtain

$$f = \frac{1}{2} \sum_{\nu} (m_{\nu}^2 + l_{\nu}^2) + \frac{\alpha}{2\beta} \log(1 - \chi) + \frac{\alpha}{2\beta} \log(1 - \phi) + \frac{\alpha}{2\beta} \frac{\chi}{1 - \chi} + \frac{\alpha}{2\beta} \frac{\phi}{1 - \phi} + \frac{\alpha}{2} \frac{Aq_1\chi}{(1 - \chi)^2} + \frac{\alpha}{2} \frac{Bp_1\phi}{(1 - \phi)^2} - \frac{1}{\beta} \left\langle \int DsDt \log \text{Tr}_{\sigma} \exp(\beta \tilde{H}) \right\rangle_{\{\xi^{\nu}\}} \quad (24)$$

with  $\nu$  denoting the condensed patterns and

$$\tilde{H} = A\sigma \left[ \sum_{\nu} m_{\nu} \xi^{\nu} + \sqrt{\alpha r} s \right] + B\sigma^2 \left[ \sum_{\nu} l_{\nu} \eta^{\nu} + \sqrt{\alpha u} t \right] + \frac{\sigma^2 \alpha A \chi}{2(1 - \chi)} + \frac{\sigma^4 \alpha B \phi}{2(1 - \phi)} \quad (25)$$

and  $A = 1/a$ ,  $B = 1/(a(1 - a))$ ,  $Ds$  and  $Dt$  Gaussian measures,  $Ds = ds(2\pi)^{-1/2} \exp(-s^2/2)$ , and

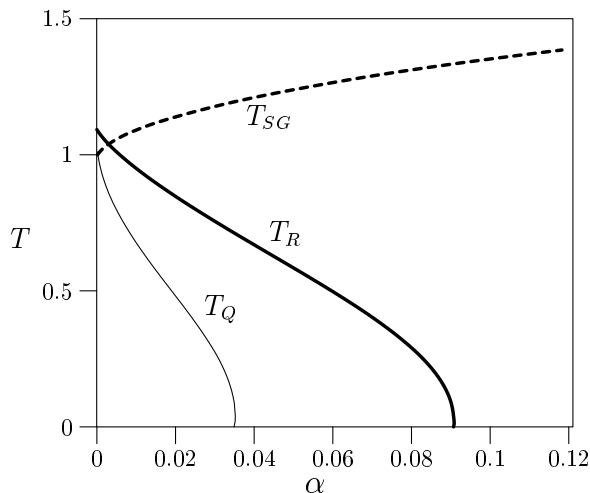
$$\chi = A\beta(q_0 - q_1), \quad \phi = B\beta(p_0 - p_1), \quad r = \frac{q_1}{(1 - \chi)^2}, \quad u = \frac{p_1}{(1 - \phi)^2}. \quad (26)$$

For  $Q = 3$  ( $\sigma^2 = \sigma^4$ ), the order parameters are defined as follows

$$\begin{aligned} m_{\nu} &= A \left\langle \xi^{\nu} \int DsDt \langle \sigma \rangle_{\beta} \right\rangle_{\{\xi^{\nu}\}} & q_1 &= \left\langle \int DsDt \langle \sigma \rangle_{\beta}^2 \right\rangle_{\{\xi^{\nu}\}} \\ l_{\nu} &= B \left\langle \eta^{\nu} \int DsDt \langle \sigma^2 \rangle_{\beta} \right\rangle_{\{\xi^{\nu}\}} & p_1 &= \left\langle \int DsDt \langle \sigma^2 \rangle_{\beta}^2 \right\rangle_{\{\xi^{\nu}\}} \\ q_0 &= p_0 = \left\langle \int DsDt \langle \sigma^2 \rangle_{\beta} \right\rangle_{\{\xi^{\nu}\}} \end{aligned} \quad (27)$$

with the small brackets  $\langle \dots \rangle_{\beta}$  denoting the usual thermal average. We recall that  $m_{\nu}$  is the overlap,  $l_{\nu}$  is related to the activity overlap,  $q_0$  is the activity of the neurons and  $q_1$  and  $p_1$  are Edwards-Anderson parameters. For one condensed pattern the index  $\nu$  can be dropped.

Solving the fixed-point equations for the order parameters and considering uniform patterns ( $a = 2/3$ ), we obtain a rich  $T - \alpha$  phase diagram (see [7] for more details). The phases that are important from a neural network point of view are presented in figure 1. The border of the retrieval phase ( $m > 0$ ,  $l > 0$ ) is denoted by a thick full line. The most important result is that the capacity of the BEG neural network is much larger than that of other  $Q = 3$  models. Compared with the  $Q = 3$ -Ising model [12], e.g., it is almost twice at  $T = 0$ . Of course this is due to the second term in the hamiltonian (22). A study of the dynamics of this model, which is in progress, confirms this result. Another new feature in the phase diagram, compared with other models, is the so-called quadrupolar phase ( $m = 0$  but  $l > 0$ ) which lies below the thin full



**Figure 1.**  $Q = 3$   $T - \alpha$  phase diagram for uniform patterns. The meaning of the lines is explained in the text.

line. It is present in the original BEG spin-model [10] and has also been seen for the extremely diluted network model [4]. In this phase the active neurons ( $\pm 1$ ) coincide with the active patterns but the sign does not. This means that although the system does not succeed in retrieval the information content is nonzero. For  $a = 2/3$  this phase lies completely within the retrieval phase but for other values of  $a$  (e.g.,  $a = 0.8$ ) it does not [7]. Besides these phases one also has a spin-glass phase and a paramagnetic phase (separated by the broken line in figure 1). The latter coexists with the retrieval phase in a region near the  $T$ -axis. We refer to [7] for further details.

In conclusion, we have presented a method starting from the mutual information between the neurons and the patterns to derive an optimal hamiltonian for a general  $Q$ -state neural network. The derivation assumes that the correlations between the neurons and patterns are small initially, and thus guarantees optimal retrieval properties (loading capacity, basin of attraction) for the model. For  $Q = 2$ , we find back the Hopfield hamiltonian for biased patterns, while for  $Q = 3$  we find the Blume-Emery-Griffiths hamiltonian. We have derived the phase diagram for this fully connected BEG model confirming that the capacity is larger than the one for related models. We believe that similar results can be obtained for vector models and other architectures. An extended version of the work on the BEG fully connected neural network will appear in [7].

## Acknowledgments

We would like to thank D. Dominguez and G.M. Shim for useful discussions. This work was supported in part by the Fund for Scientific Research, Flanders (Belgium).

## References

- [1] Dominguez D R C and Bollé D 1998 *Phys. Rev. Lett.* **80** 2961
- [2] Bollé D and Dominguez Carreta D 2000 *Physica A* **286** 401
- [3] Amit D J, Gutfreund H G and Sompolinsky H 1987 *Phys. Rev. A* **35** 2293
- [4] Carreta Dominguez D and Korutcheva E 2000 *Phys. Rev E* **62** 2620
- [5] Blahut R E 1990 *Principles and Practice of Information Theory*, (Addison-Wesley, Reading, MA) Chapter 5.
- [6] Hopfield J J 1982 *Proc. Nat. Acad. Sci. USA* **79** 2554
- [7] Bollé D and Verbeiren T, in preparation
- [8] de Almeida R M C and Iglesias J R 1990 *Phys. Lett A* **146** 239
- [9] Bollé D, Huyghebaert J and Shim G M 1994 *J. Phys A: Math. Gen.* **27** 5871
- [10] Blume M, Emery J and Griffiths R B 1971 *Phys. Rev. A* **4** 1071
- [11] Mézard M, Parisi G and Virasoro M A 1978 *Spin Glass Theory and Beyond* (Singapore: World Scientific)
- [12] Bollé D, Rieger H and Shim G M 1994 *J. Phys. A: Math. Gen.* **27** 3411